Common knowledge of rationality in games

de Bruin, B.P.

Published in:
Notre Dame Journal of Formal Logic

DOI:
10.1215/00294527-2008-011

Publication date:
2008

Notre Dame Journal of Formal Logic
Volume 49, Number 3, 2008

Common Knowledge of Rationality in Extensive Games

Boudewijn de Bruin

Abstract We develop a logical system that captures two different interpretations of what extensive games model, and we apply this to a long-standing debate in game theory between those who defend the claim that common knowledge of rationality leads to backward induction or subgame perfect (Nash) equilibria and those who reject this claim. We show that a defense of the claim à la Aumann (1995) rests on a conception of extensive game playing as a one-shot event in combination with a principle of rationality that is incompatible with it, while a rejection of the claim à la Reny (1988) assumes a temporally extended, many-moment interpretation of extensive games in combination with implausible belief revision policies. In addition, the logical system provides an original inductive and implicit axiomatization of rationality in extensive games based on relations of dominance rather than the usual direct axiomatization of rationality as maximization of expected utility.

1 Introduction

There is wide disagreement between game theorists and logicians about the epistemic foundations of the solution concept of backward induction or subgame perfect (Nash) equilibrium. The disagreement centers on the question whether or not common knowledge of rationality leads to backward induction. Aumann [2; 3; 4] argues in favor of this claim, while related but weaker claims have been defended by Broome and Rabinowicz [16], and Rabinowicz [33]. Reny [34; 35] argues against the claim, and related claims have been defended by Basu [8], Ben-Porath [9], Bicchieri [11], Binmore [12], Clausing [20], and Stalnaker [39].

The purpose of this paper is not to substantiate one line of argument or another. Rather, by analyzing the logical form of the arguments à la Aumann [2] and à la Reny [34], we point out that important modeling assumptions have been overlooked. We have devised a logical formalism that captures two different interpretations of

Received January 18, 2007; accepted January 11, 2008; printed June 6, 2008
2000 Mathematics Subject Classification: Primary, 03B42, 91A18, 91A26
Keywords: common knowledge, epistemic characterization theorem, rationality
 10.1215/00294527-2008-011

what extensive games in fact model: they are models of one-shot forms of strategic interaction or of many-moment forms of interaction.

The first main claim of this paper is that the argument a la Aumann [2] assumes the one-shot interpretation in combination with a principle of rationality that is incompatible with it. The second main claim is that the argument à la Reny [34] assumes the many-moment interpretation in combination with implausible belief revision policies.

From a game theoretic point of view, the originality of the approach outlined here inheres in its espousal of two interpretations of extensive game play, in the formalization of these interpretations and in laying bare assumptions the relevance of which to arguments about backward induction and common knowledge has hitherto gone unnoticed.

From a logical point of view, the originality of our approach lies in offering a new way to formalize game theoretic rationality. Whereas in our formalization of the argument à la Reny [34] we adopt the standard treatment in terms of expected utility maximization, in our formalization of the argument à la Aumann [2] we present an inductive and implicit axiomatization of rationality. To our knowledge, this is the first time this has been done.

To put it slightly imprecisely, the thrust of this paper is not so much to prove new theorems but rather to disentangle new assumptions underlying old results. We show the logician how to put to use his or her formalism in other fields, and we show the game theorist how to shed light on his or her formalism with greater degrees of precision.

An important earlier comparative study is that of Halpern [26]. The difference from the present analysis is that Halpern compares Aumann [2] with Stalnaker [38] rather than with Reny [34]. Moreover, Halpern does not distinguish between one-shot and many-moment interpretations of extensive game play, and he does not give an inductive and implicit axiomatization of rationality. Relevant general references in the logic and games literature include Aumann [5], Baltag [6], van Benthem [10], Bonanno [13; 14], Feinberg [23; 24], Kaneko [27], Pauly [30; 31], Pietarinen [32], and Wolter [41].

Section 2 of this paper gives a brief survey of game theoretic and logical notational conventions (some familiarity with game theory and epistemic logic is assumed). Section 3 deals with the one-shot interpretation, while Section 4 treats the many-moment interpretation. Both sections first introduce the interpretation before going on to explore the necessary axioms. The relevant theorem is stated and proved, and, finally, we turn to a discussion. Conclusions are presented in Section 5.

2 Notation

2.1 Game theory An *extensive (form) game* Γ with *perfect information* and players from I is based on a finite tree $(X, \prec, \rho)$ where ρ is the root or starting point of the game and $\prec$ a strict (irreflexive and transitive) partial ordering of the nodes such that $\rho \prec x$ for all $x \neq \rho$. The inverse is written $\succ$. Nodes x without $y \succ x$ are called *terminal* nodes.

The *depth* $d(x)$ of a node x is roughly the maximum number of edges connecting x with a terminal node; roughly, for an inductive definition, would be more precise.

A *player function* ι associates all nonterminal or *decision* nodes D with elements from I indicating which nodes are within a player's control.

Utility functions $u_i : X \setminus D \to \mathbb{R}$ are used to represent the players' von Neumann and Morgenstern preference orderings. An extensive game Γ is called *generic* whenever all u_i are injective.

A *strategy* of player i is a function mapping all his decision nodes to immediate successors. That is, a function $s : \iota^{-1}(i) \to X$ such that $x \prec s(x)$ but $x \prec y \prec s(x)$ for no y. In general, we use the term *strategy* for the function s on $\iota^{-1}(i)$ (a function defined on all decision nodes of player i) and restrict the use of the term *action* to the restriction of s to one single decision node of player i.

The *subgame* generated by x is simply the game based on the tree generated by x with the obvious restrictions for the player function and the utility functions. If Γ is an extensive game, the subgame generated by some decision node x is written Γ_x.

The *normal form* $nf(\Gamma)$ of an extensive game Γ is a triple $(I, (S_i)_i, (v_i)_i)$ where I collects the set of players of Γ, S_i all strategies player i has in Γ, and $v_i : \prod_i S_i \to \mathbb{R}$ are utility functions such that

$$v_i(s_1, \ldots, s_i, \ldots, s_N) = u_i(O(s_1, \ldots, s_i, \ldots, s_N)),$$

where O is a function mapping a tuple of strategies to the terminal node of the extensive game that is reached when the players play according to these strategies. We shall write $u_i(s, t)$ for $v_i(O(s, t))$. If $nf(\Gamma) = (I, (S_i)_i, (v_i)_i)$, and if X_1, X_2, and so on, are sets of strategies satisfying $X_i \subseteq S_i$ for all $i \in I$, then the *subspan* of $nf(\Gamma)$ with respect to $\prod_i X_i$ is the triple $(I, (X_i)_i, (v_i|_{X_i})_i)$ obtained from $nf(\Gamma)$ by removing for all i the strategies in the complement of X_i (with respect to S_i) and modifying the utility functions correspondingly.

We write $\mathsf{nsd}_i^{\Gamma}(X_1, \ldots, X_N)$ for the strategies that are not strictly dominated (in other words, strictly *undominated*) for player i in the subspan of $nf(\Gamma)$ with respect to $\prod_i X_i$, that is, strategies for which there is no strategy in X_i which does strictly better against any combination of opponents' strategies. We are in general interested in dominance relations in subspans of the normal form of subgames of some underlying game, that is, in constructs of the form $\mathsf{nsd}_i^x(X_1, \ldots, X_N)$, where $X_i \subseteq S_i$. To compute such sets (the idea is simpler than the construction), consider the subgame Γ_x of Γ generated by x, construct its normal form $nf(\Gamma_x)$, delete from $nf(\Gamma_x)$, for all j, the strategies not coinciding on Γ_x with any strategy from X_j, find out which of the remaining strategies in the resulting subspan of $nf(\Gamma_x)$ are strictly undominated, and then take all strategies from S_i coinciding on Γ_x with such a strictly undominated strategy. For weak dominance, we define nwd_i and its relativizations similarly.

The set containing the strategies coinciding with the *backward induction* strategy on the subgame generated by x is written BI_i^x. Assuming that the extensive game is *generic* in the sense that no player is indifferent between any two terminal nodes, BI_i^x contains all strategies prescribing the uniquely optimal action at x if x is an immediate predecessor of a terminal node. Reasoning downward to the root of the game, BI_i^z collects all strategies prescribing the uniquely optimal action at decision node z under the assumption that at decision nodes $y \succ z$ higher up in the game tree i plays according to BI^y. Clearly, BI_i^{ρ} is a singleton also written BI_i. For terminal nodes x we use the convention that $\mathsf{BI}_i^x = S_i$.

2.2 Logic The logical symbols used are negation and the connectives (where $\bigwedge$ $(\bigvee)$ is used for large conjunctions (disjunctions)), epistemic operators $\mathbf{K}_i$ (knowledge of player i), $\mathbf{E}_I$ (knowledge of every player $i \in I$), and $\mathbf{C}_I$ (common knowledge

among all players $i \in I$). We often write X for $\bigvee X$ whenever X is a set of propositional formulas for strategies, using the term *propositional formula* throughout as a relatively neutral term lacking the connotation of atomic proposition or prime formula that the term *proposition letter* may carry (see Barwise [7], p. 23).

Epistemic operators with superscript x for decision node x appear in the many-moment interpretation, as do probabilistic doxastic operators $\mathbf{P}_i^x(\varphi) = p_k$, for knowledge and probabilistic beliefs of player i at the decision moment at which decision node x is reached, with, in principle, uncountably many symbols p_k.

Nonlogical symbols include propositional formulas i_k (player i plays his kth strategy), i_k^x (player i plays according to his kth strategy in the subgame generated by x), and $i_k(x)$ (player i plays, at x, the action prescribed by his kth strategy).

When writing about utility we use constructs of the form $u_i(k, l) = r$ (the utility to player i of playing his kth strategy against opponent strategy l is r), with, in principle, uncountably many symbols r, and $u_i^x(k, l) = r$ (the utility to player i of playing his kth strategy in the subgame generated by x against opponent strategy l is r). For our purposes it is entirely harmless to assume an uncountable language.

For various rationality principles we use propositional formulas Nrat_i^x (player i is rational, in the subgame generated by x, at all reached decision nodes; *on-path* rationality, see below), Frat_i (player i is rational, in all subgames of the game and at all decision nodes, reached or unreached; *off-path* rationality, or rationality in Aumann's sense, see below), and Rrat_i^x (player i is rational, in the subgame generated by x, in Reny's sense), with primed variants.

The prime formulas p of the formalism are, then, of the forms i_k, i_k^x, $i_k(x)$, $u_i(k, l) = r$, $u_i^x(k, l) = r$, Nrat_i^x, Frat_i, and Rrat_i^x, and the well-formed formulas are derived from them in the usual manner:

$$\varphi :: = p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathbf{K}_i\varphi \mid \mathbf{E}_I\varphi \mid \mathbf{C}_I\varphi \mid \mathbf{P}_i(\varphi) = q.$$

We use $\mathsf{nsd}_i(X_1, \ldots, X_N)$, $\mathsf{nwd}_i(X_1, \ldots, X_N)$, and Bl_i, and superscripted variants for subgames, for the sets of the obvious propositional formulas of strategies (and, with the above convention, for the disjunctions of their elements).

Standard doxastic and epistemic axioms are used with usual abbreviations K, T, D, 4, 5. For common knowledge, the C axiom is $\mathbf{C}_I\varphi \leftrightarrow \mathbf{E}_I(\varphi \wedge \mathbf{C}_I\varphi)$. Standard axioms for linear (in)equalities, axioms fixing the Kolmogorov interpretation of probability theory, and axioms inter-relating probabilistic and nonprobabilistic beliefs are needed (Fagin and Halpern [22]). In particular, we need the following two axioms:

Cons	$\Box_i\varphi \leftrightarrow \mathbf{P}_i(\varphi) = 1.$
KnProb	$\mathbf{P}_i(\varphi) = q \rightarrow \Box_i\mathbf{P}_i(\varphi) = q.$

The proof rules of modus ponens and necessitation are standard. For common knowledge, the rule of induction is

$$\text{If } \vdash \varphi \rightarrow \bigwedge_{i \in I} \mathbf{K}_i(\varphi \wedge \psi), \text{ then } \vdash \varphi \rightarrow \mathbf{C}_I\psi.$$

3 One-Shot Interpretation

The mathematical differences between normal form games and extensive games strongly suggest that the former model situations of simultaneous and independent choice while the latter model temporally extended situations of sequential choice.

But are we forced to adopt such a perspective? The one-shot interpretation agrees with the founding fathers of game theory in answering this question negatively: playing an extensive game is playing its normal form, because extensive form and normal form are "strictly equivalent" (von Neumann and Morgenstern [40], p. 85). That is, whenever players play an extensive game, what they really do is choose, at one point in time, their strategies for the entire game (Fudenberg and Tirole [25], p. 85; Osborne and Rubinstein [29], pp. 94–95).

This does not completely determine a unique one-shot interpretation though. There is room for disagreement about the relevant kind of rationality principles, for one could decide to invoke some aspects of the sequential structure of the game to ascertain whether a strategy is rational or not. A strategy maps all decision nodes of a player to actions. But choosing, for some decision node, one action over another implies that certain decision nodes will not be reached. Or, in a vocabulary that is more in line with the one-shot interpretation, the action prescribed by the strategy at such decision nodes will not influence the outcome of the game.

To capture the differences, a strategy is called *on-path* rational whenever the rationality depends only on what happens on the actual path through the extensive game; it is called *off-path* rational if it prescribes rational actions at every decision node, reached or unreached.

Aumann [2] asserts that

> each player chooses a *strategy*, in the usual game theoretic sense of the term...; that is, he decides what to do at each of his vertices x in the game tree, whether or not x is reached. (Aumann [2], p. 7, notation changed)

This seems to demonstrate that he adopts a one-shot conception of extensive game play. Yet he also writes that "when deciding what to do at x, the player considers the situation *from that point on:* he acts *as if* x is reached" (Aumann [2], p. 7, emphasis his, notation changed). He concludes that

> it is this feature that distinguishes the current analysis from a strategic [i.e., normal] form analysis. (Aumann [2], p. 7)

If we attribute to Aumann the one-shot interpretation and also accept this conclusion, then, contrary to what we suggested earlier, there seems to be a difference between the one-shot interpretation and playing the normal form of an extensive game. But there is no inconsistency. When we define the one-shot interpretation in terms of normal form game play, we focus on the objects of choice. According to the one-shot interpretation, the objects of choice of an extensive game are in fact the strategies of its normal form.

And Aumann agrees. When he goes on to contrast his view with the strategic form analysis he is in fact concerned with rationality principles, not with objects of choice. For Aumann, to evaluate the rationality of a strategy one has to go beyond the information of the normal form and inspect the prescriptions of the strategy at all decision nodes of the underlying extensive game. That is, he adopts a one-shot interpretation of extensive game play with an off-path conception of rationality. This is underscored by his statement that a rational player,

> no matter where he finds himself—at which vertex—[,]... will not knowingly continue with a strategy that yields him less than he could have gotten with a different strategy. (Aumann [2], p. 7)

Or, in alternative wording,

> For each of his vertices x and strategies k, it is not the case that [player] i knows that k would yield him a higher conditional payoff at x than the strategy he chooses. (Aumann [2], p. 10, notation changed)

All in all, Aumann adopts a one-shot interpretation with off-path rationality. This is the same as playing normal form games as far as the objects of choice are concerned. But it is different with respect to the rationality principle.

3.1 Axioms To capture the one-shot interpretation in formal terms, the following axioms are used, given an extensive game with perfect information Γ:

$\text{Strat}_{\geq 1}$	$\bigwedge_i \bigvee_k i_k$.
$\text{Strat}_{\leq 1}$	$\bigwedge_i \bigwedge_{k \neq l} \neg(i_k \wedge i_l)$.
KnStrat	$\bigwedge_i \bigwedge_k (i_k \leftrightarrow \mathbf{K}_i i_k)$.
Sub_1	$\bigwedge_i \bigwedge_k (i_k^x \leftrightarrow \bigvee_{l \in D} i_l)$ where D contains the indices of the strategies coinciding with k on the subgame generated by x.
Sub_2	$\bigwedge_i \bigwedge_k (i_k(x) \leftrightarrow \bigvee_{l \in D} i_l)$ where D contains the indices of the strategies coinciding with k on decision node x.
UtSub	$\bigwedge_i \bigwedge_{k,l,m,n} (u_i^x(k,m) = u_i^x(l,n))$ whenever i's kth and lth, and j's mth and nth, strategies coincide on the subgame generated by x.

The first three axioms establish that players have to pick exactly one strategy and that they have to know what they do. Sub_1 states that the use of a superscript does indeed involve the restriction of some strategy to the relevant subgame; Sub_2, that function notation is used to talk about the action taken at some decision node; UtSub, that the superscript works well when applied to utility functions. Although these axioms are not very interesting in themselves, they are necessary to determine what it is to play a game. Without them, players could refuse to act or choose more than one strategy or act unknowingly and so on. That is, these axioms are simply there to fix the meaning of some of the propositional formulas of our formal system.

Similarly fixing the meaning of propositional formulas, but more interesting in itself, is our way of formalizing on-path and off-path rationality. The originality of our approach is not to axiomatize rationality in terms of expected utility maximization, but rather inductively and implicitly by means of the following three axioms, phrased, without loss of generality, for two players i and $j \neq i$:

NRat_{bas}	$\mathsf{Nrat}_i^x \rightarrow \mathsf{nsd}_i^x(S_i, S_j)$.
NRat_{ind}	$(\mathsf{Nrat}_i^x \wedge \mathbf{K}_i X_i \wedge \mathbf{K}_i X_j) \rightarrow \mathsf{nsd}_i^x(X_i, X_j)$.
FRat	$\mathsf{Frat}_i \leftrightarrow \bigwedge_{\rho \preceq x} \mathsf{Nrat}_i^x$.

These axioms need some explanation. First a preliminary remark about applying on-path rationality to subgames: In the one-shot interpretation, objects of choice, strategies, are always functions mapping all decision nodes of a player to actions. Beliefs, then, are beliefs about which such strategies opponents will choose. However, it makes perfect sense to speak about the on-path rationality of a strategy in any subgame, for one can consider the restriction (to the subgame) of a strategy and evaluate its rationality as a course of action in the subgame in the light of the restrictions (to the same subgame) of the strategies one expects one's opponents to play. It is this idea that is captured in the first two axioms.

What it is rational to do often depends essentially on one's beliefs, but not always. The first axiom captures the base case without beliefs. It states that player i, if on-path rational in the subgame Γ_x generated by x, will never choose a strategy which prescribes bad actions independently of what his opponents play: if player i is on-path rational in Γ_x, he will not choose any strategy of which the restriction to Γ_x coincides with a strategy strictly dominated in the normal form of Γ_x.

The second axiom states that, if he is on-path rational in Γ_x, player i will never play a strategy that is strictly dominated in the normal form of Γ_x from which those strategies (both of his opponents as well as himself) have been removed that he believes will not be chosen. The beliefs are represented by sets X_i and X_j of strategies. The third axiom, finally, states that player i is off-path rational in the entire game if he is on-path rational in all of its subgames.

Before turning to the characterization result, we emphasize some of the features of our inductive and implicit axiomatization of rationality. First, the axioms give necessary conditions for rationality, but no sufficient conditions. While this may seem a technical drawback of the axiomatization, there is in fact a conceptual reason why it is impossible to phrase sufficient conditions in terms of the behavior of the players only as players may be imagined who stumble upon rational actions by accident rather than by deliberation. That a player obtains the best possible score in a game, for instance, does not entail he played rationally.

Second, in contrast to the standard way of defining rationality in terms of expected utility maximization (the route followed in the section on Reny), our approach makes it easy to reveal procedural aspects of epistemic characterization results. Drawing a line between a base case without beliefs and an inductive step with beliefs makes it possible to mimic steps of removing noninductive strategies by steps in the hierarchy of common knowledge. This becomes very explicit in the inductive character of the proof of the characterization result.

Third, there is an interesting link to findings from experimental economics (or behavioral game theory, to be precise). Many experiments suggest that actual players only converge to playing backward induction over time (Camerer [19], and references therein). An obvious interpretation in terms of our formalism is that it takes some time for the players to become aware of full inductive reasoning and higher levels of common knowledge. In other words, they need their time to prove the characterization result.

Fourth, as we shall see, the present inductive and implicit axiomatization makes it easy to study variant rationality notions in alternative characterization results.

3.2 Characterization result Given an extensive game with perfect information Γ, let proof system ${}_{\Gamma}\mathbf{K_C}\mathsf{Frat}$ consist of the following axioms: all propositional tautologies, K, C, the proof rules modus ponens, necessitation, and induction, all axioms for one-shot game-playing situations for Γ, plus the three rationality axioms presented above. The claim à la Aumann that common knowledge of rationality leads to backward induction is the following theorem. In fact, since our rendering does not use the T axiom, the epistemic assumption is better described as common true belief about rationality.

Theorem 3.1 *Let Γ be a finite generic N-person extensive game with perfect information. Then, for all decision nodes x,*

$$\vdash_{\Gamma \mathbf{K_C}\mathsf{Frat}} (\mathbf{C}\mathsf{CFrat} \wedge \mathsf{Frat}) \rightarrow \bigwedge_i \mathsf{BI}_i^x.$$

To prove the theorem we need two lemmas. First, to collect all (propositional formulas for) backward induction strategies in any subgame Γ_x, we take all strategies that prescribe backward induction actions in all real subgames of Γ_x and obtain the set $\bigcap_{x \prec y} \mathsf{BI}_i^y$. Some of the strategies in this set, however, do not prescribe the backward induction action at x, and therefore we restrict attention to those elements for which there is no strictly better alternative given that all players take backward induction actions at decision nodes $y \succ x$. This is establishes the first lemma.

Lemma 3.2

$$\mathsf{BI}_i^x = \mathsf{nsd}_i^x\Big(\bigcap_{x \prec y} \mathsf{BI}_1^y, \ldots, \bigcap_{x \prec y} \mathsf{BI}_N^y\Big) \cap \bigcap_{x \prec y} \mathsf{BI}_i^y.$$

In the formalism proposed, the formula $\bigwedge_{x \prec y} \bigvee \mathsf{BI}_i^y$ states that at any $y \succ x$ player i plays according to backward induction (if y is a decision node of his). The intersection $\bigcap_{x \prec y} \mathsf{BI}_i^y$ not being empty (it contains all strategies available to i in Γ that prescribe backward induction actions in Γ_x), it is straightforward to observe that $\vdash_{\Gamma \mathbf{K_C}\mathsf{Frat}} \bigwedge_{x \prec y} \bigvee \mathsf{BI}_i^y \rightarrow \bigvee \bigcap_{x \prec y} \mathsf{BI}_i^y$. With the convention to omit disjunction symbols in front of sets of propositional formulas, this establishes the second lemma.

Lemma 3.3

$$\vdash_{\Gamma \mathbf{K_C}\mathsf{Frat}} \bigwedge_{x \prec y} \mathsf{BI}_i^y \rightarrow \bigcap_{x \prec y} \mathsf{BI}_i^y.$$

We are now ready for the proof of Theorem 3.1.

Proof We prove the result for $N = 2$ with players i and $j \neq i$. For more than two players one only needs to add the relevant conjuncts and to expand nsd_i^x to a function taking three or more arguments. Suppressing the proof system, we have $\vdash \mathsf{Frat}_i \rightarrow \mathsf{Nrat}_i^x$ by axiom FRat, and the case of $d(x) = 1$ reduces to axiom NRat_{bas} with an application of Lemma 3.2. Let $d(x) > 1$. The inductive hypothesis gives for every $y \succ x$

$$\vdash (\mathbf{C}\mathsf{CFrat} \wedge \mathsf{Frat}) \rightarrow \mathsf{BI}_i^y.$$

Because we are dealing with finite games we can aggregate the proofs for all $y \succ x$ and both players i and j into

$$\vdash (\mathbf{C}\mathsf{CFrat} \wedge \mathsf{Frat}) \rightarrow \Big(\bigwedge_{x \prec y} \mathsf{BI}_i^y \wedge \bigwedge_{x \prec y} \mathsf{BI}_j^y\Big),$$

and, applying Lemma 3.3, into

$$\vdash (\mathbf{C}\mathsf{CFrat} \wedge \mathsf{Frat}) \rightarrow \Big(\bigcap_{x \prec y} \mathsf{BI}_i^y \wedge \bigcap_{x \prec y} \mathsf{BI}_j^y\Big).$$

An application of the necessitation rule for $\mathbf{K}_i$, and the K axiom, together with some propositional reasoning, yields

$$\vdash \mathbf{C}\mathsf{CFrat} \rightarrow \Big(\mathbf{K}_i \bigcap_{x \prec y} \mathsf{BI}_i^y \wedge \mathbf{K}_i \bigcap_{x \prec y} \mathsf{BI}_j^y\Big).$$

Since, as we saw, $\vdash \mathsf{Frat}_i \rightarrow \mathsf{Nrat}_i^x$, we obtain

$$\vdash (\mathbf{C}\mathsf{Frat} \wedge \mathsf{Frat}) \rightarrow \big(\mathsf{Nrat}_i^x \wedge \mathbf{K}_i \bigcap_{x \prec y} \mathsf{Bl}_i^y \wedge \mathbf{K}_i \bigcap_{x \prec y} \mathsf{Bl}_j^y\big).$$

Applying the NRat_{ind} axiom we arrive at

$$\vdash (\mathbf{C}\mathsf{Frat} \wedge \mathsf{Frat}) \rightarrow \mathsf{nsd}_i^x\big(\bigcap_{x \prec y} \mathsf{Bl}_i^y, \bigcap_{x \prec y} \mathsf{Bl}_j^y\big).$$

Invoking the inductive hypothesis again, and applying Lemma 3.3, we can make the consequent of this formula somewhat more precise in

$$\vdash (\mathbf{C}\mathsf{Frat} \wedge \mathsf{Frat}) \rightarrow \mathsf{nsd}_i^x\big(\bigcap_{x \prec y} \mathsf{Bl}_i^y, \bigcap_{x \prec y} \mathsf{Bl}_j^y\big) \cap \bigcap_{x \prec y} \mathsf{Bl}_i^y,$$

which is

$$\vdash (\mathbf{C}\mathsf{Frat} \wedge \mathsf{Frat}) \rightarrow \mathsf{Bl}_i^x,$$

by Lemma 3.2. □

3.3 Discussion We have presented two varieties of the one-shot interpretation of extensive game play: one with on-path rationality and one with off-path rationality. Aumann was seen to adopt the latter version. We do not believe, however, that the latter version is conceptually consistent. More precisely, we believe that off-path rationality is strictly incompatible with the true spirit of the one-shot interpretation.

The reason is that there is no sensible rationale behind taking care of what would happen at unreached, off-path nodes in a situation in which the objects of choice are strategies from the normal form of an extensive game. In a one-shot situation it just does not make sense to talk about nodes being reached or not. The game-playing situation is a strategic predicament in which the players choose a strategy that fixes a complete plan of action for the entire game. Temporal deliberation is senseless, as is thinking about players having beliefs at various points in a temporally extended sequence of decision moments. No nodes are reached or unreached. There is only one decision moment and the outcome of the game is determined on the basis of the strategies the players choose at that precise decision moment.

Does the fact that the one-shot interpretation leaves no room for rationality notions that transcend the normal form entail that the epistemic characterization result of backward induction fails to be significant, or that backward induction cannot be epistemically characterized in a one-shot interpretation? We answer the first question affirmatively: there is no sense to any epistemic characterization that presupposes the one-shot interpretation together with a form of rationality that goes beyond on-path rationality by using the specific structural properties of extensive games.

The second question, however, need not be answered affirmatively. It is not difficult to see that once we rephrase the NRat axioms in terms of weak rather than strict dominance, backward induction can be characterized on the basis of on-path rationality at the root of the game only. Given an extensive game with perfect information Γ, let proof system ${}_\Gamma\mathbf{K}_\mathbf{C}\mathrm{Nrat}'$ consist of the following axioms: all propositional tautologies, K, C, the proof rules modus ponens, necessitation, and induction, all axioms for one-shot game-playing situations for Γ, plus the following two rationality axioms:

NRat'_{bas} $\quad \mathsf{Nrat}'^x_i \rightarrow \mathsf{nwd}_i^x(S_i, S_j)$.

NRat'_{ind} $\quad (\mathsf{Nrat}'^x_i \wedge \mathbf{K}_i X_i \wedge \mathbf{K}_i X_j) \rightarrow \mathsf{nwd}_i^x(X_i, X_j)$.

The claim that common knowledge of rationality in terms of weak dominance leads to backward induction is the following theorem.

Theorem 3.4 *Let Γ be a finite generic N-person extensive game with perfect information. Then, for all decision nodes x,*

$$\vdash_{\Gamma \mathbf{K_C} \mathsf{Nrat}'} (\mathbf{C}\mathsf{Nrat}'^{\rho} \wedge \mathsf{Nrat}'^{\rho}) \rightarrow \bigwedge_i \mathsf{BI}_i^{\rho}.$$

To prove this theorem, first observe that on the level of the normal form of the extensive game, the relevant solution concept is iterated weak dominance. The actual outcome of a process of iterative elimination of weakly dominated strategies depends on the exact definition of the elimination algorithm (Fudenberg and Tirole [25], p. 461). For our purposes, however, a lemma due to Moulin [28] shows this to be irrelevant.

Lemma 3.5 *Let Γ be a finite generic N-person extensive game with perfect information, and let $nf(\Gamma)$ be its normal form. Then*

1. *the natural algorithms for iterated weak dominance yield a unique strategy profile in $nf(\Gamma)$;*
2. *these algorithms all yield the same strategy profile;*
3. *the strategies from this profile correspond to the backward induction strategies of Γ.*

The proof of Theorem 3.4 is, then, a straightforward analogue of the proof of Theorem 3.1.

This is all very well, but it precipitates us into another kind of problem, for the rationality of weak dominance, as well as the solution concept of iterated weak dominance, are not unproblematic (Asheim and Dufwenberg [1]; Brandenburger, Friedenberg, and Keisler [15]; Samuelson [36]). A discussion of the problems falls outside the scope of this paper. References to alternative attempts to characterize backward induction in terms of common knowledge are given in Section 1.

Before proceeding, we should mention that we have left undiscussed and unformalized the background assumption that the utility functions of the players are commonly known among them. It is clear that this assumption is necessary for the characterization result. If, for instance, player i does not know that player j knows i's utility structure, player i cannot from knowledge of j's knowledge about i's rationality deduce anything about j's knowledge about i's prospective strategy choice.

It would be fairly easy to formalize common knowledge of utility though. An axiom scheme such as

$$\text{KnUt} \qquad \bigwedge_i \bigwedge_{k,l} (u_i(k,l) = r \rightarrow \mathbf{K}_i u_i(k,l) = r),$$

would capture the fact that player i knows what utility he attaches to any terminal node. Ranging over an (if you wish, finite) set of real numbers including the correct utilities (that is, including the set $\{x | u_i(k,l) = x \text{ for some } k,l,i\}$), the antecedent $u_i(k,l) = r$ turns out true for the utility r that i assigns to $O(k,l)$, while the consequent says that i knows he so assigns utility. In fact, in a critique of the epistemic characterization of the solution concept studied by Dekel and Fudenberg [21] we have formalized common knowledge of the fact that players are approximately correctly informed about their opponents' utility functions (de Bruin [18]).

For present purposes, however, no analytical clarity would be gained by making common knowledge of utility explicit in the formalism. The logically and game

theoretically interesting distinctions here have to do with rationality, not with utility. That is why we have left common knowledge about utility unformalized.

4 Many-Moment Interpretation

The one-shot interpretation of extensive games deals conceptually with situations in which players choose a strategy for a sequential strategic decision problem at one point in time; there is one and only one decision moment at which a strategy for the entire game is chosen. The many-moment interpretation, by contrast, takes extensive games as models of a temporal succession of many decision moments; the objects of choice are actions at decision nodes, not strategies for the entire game. A different view of the objects of choice brings with it a different view of what the players' beliefs and rationality principles are. At every decision moment a player has beliefs about the future development of the game, which, in principle, may change over time; moreover, they range over actions, not strategies. Similarly, rationality principles pertain to actions rather than strategies, and they may change over time, too.

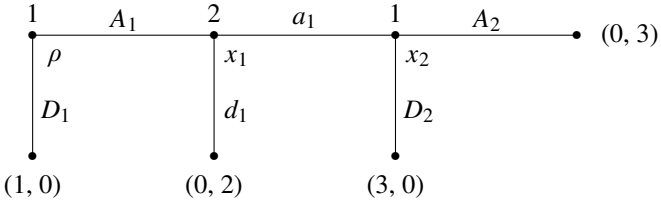

Figure 1 Reny's Game

Reny [34] clearly adopts the many-moment interpretation, referring to the game shown in Figure 1:

> I claim that if player one does not take the dollar and end the game in the first round [does not play D_1], but instead leaves it so that player 2 must decide whether or not to take the two dollars [whether or not to play d_1], then it is no longer possible for rationality to be common knowledge. (i.e. At [*sic*] player two's information set, it is not possible for rationality to be common knowledge). (Reny [34], pp. 364–65)

Such reasoning makes no sense in the one-shot interpretation, according to which no decision nodes are reached at all. On the contrary, strategies are chosen which may or may not induce a path through the game tree to reach some decision node. But it does not make sense to speak about the beliefs or the knowledge of the players at those decision nodes. The players have beliefs at the moment they choose their strategy, but the game stops after that.

Reny, by contrast, considers beliefs of a player at some decision moment. Such beliefs describe the expectations of a player about what actions will be taken at all decision nodes in the subgame generated by the current decision node. In principle, the many-moment interpretation would leave two possibilities open. The beliefs at some decision moment could, first, be viewed as dependent on what happened before; that is, they would be sensitive to the history of the decision moment in the

sense, for instance, that a player could decide to believe his opponent to be irrational if the current decision node can only be reached by the irrational play of his opponent. Second, the beliefs could be completely insensitive to the history. We shall see that Reny's inconsistency result presupposes a history-sensitive view of belief formation.

4.1 Axioms Before presenting the axioms for the many-moment interpretation we need to set down some notation. Without loss of generality, we consider a two-person extensive game with players i and $j \neq i$. Given beliefs $\mathbf{P}_i^x(i_k^x) = p_k$ and $\mathbf{P}_i^x(j_l^x) = p_l$, first define an auxiliary and admittedly quite strange notion of something like the expected utility conditional on reaching some immediate successor $y \succ x$ as

$$\mathrm{EU}_i(y, \mathbf{P}_i^x) = \sum_{k,l} p_k p_l u_i^y(k, l).$$

Then define $\mathrm{EU}_i(k, x, \mathbf{P}_i^x)$, the intended interpretation being the expected utility of playing according to the kth strategy at the decision moment at which node x is reached, as

$$\mathrm{EU}_i(k, x, \mathbf{P}_i^x) = \mathrm{EU}_i(y, \mathbf{P}_i^x)$$

for that y that is reached when at x player i plays according to his kth strategy.

To capture the many-moment interpretation formally, the following axioms are used, given an extensive game with perfect information Γ:

$\mathrm{Strat}_{\geq 1}$	$\bigwedge_i \bigvee_k i_k$.
$\mathrm{Strat}_{\leq 1}$	$\bigwedge_i \bigwedge_{k \neq l} \neg(i_k \wedge i_l)$.
Sub_1	$\bigwedge_i \bigwedge_k (i_k^x \leftrightarrow \bigvee_{l \in D} i_l)$ where D contains the indices of the strategies coinciding with k on the subgame generated by x.
Sub_2	$\bigwedge_i \bigwedge_k (i_k(x) \leftrightarrow \bigvee_{l \in D} i_l)$ where D contains the indices of the strategies coinciding with k on decision node x.
UtSub	$\bigwedge_i \bigwedge_{k,l,m,n} (u_i^x(k, m) = u_i^x(l, n))$ whenever i's kth and lth, and j's mth and nth strategies coincide on the subgame generated by x.
KnStratM	$\bigwedge_i \bigwedge_k (i_k \leftrightarrow \bigwedge_{\rho \preceq x} \mathbf{K}_i^x i_k)$.
KnWhere	$\bigwedge_i \mathbf{K}_i^x \bigwedge_j \bigvee_{y \prec x, k \in D} j_k^y$ where D contains the indices of the strategies that are consistent with reaching x.

The first five axioms are those axioms from the one-shot interpretation that do not contain a $\mathbf{K}_i$. The motivation is similar. Let us now turn to the last two axioms. KnStratM ensures that at every moment during a game, players know what their choice of strategy is. This is a strong assumption because it entails that they know, too, what they will choose at every possible future occasion. KnWhere is there to guarantee that players know, at some decision moment, which decision node has been reached.

To formalize rationality we need one axiom:

RRat $\quad \mathsf{Rrat}_i^x \leftrightarrow$

$$\begin{aligned} &((\mathbf{K}_i^x \textstyle\bigwedge_{k,l} u_i^x(k, l) = r_{i,k,l} \wedge \\ &\textstyle\bigwedge_k \mathbf{P}_i^x(i_k^x) = p_k \wedge \\ &\textstyle\bigwedge_l \mathbf{P}_i^x(j_l^x) = p_l \wedge \end{aligned}$$

$$i_m(x)) \rightarrow \bigwedge_k \mathrm{EU}_i(m, x, \mathbf{P}_i^x) \geq \mathrm{EU}_i(k, x, \mathbf{P}_i^x)).$$

This is a straightforward relativization to subgames of the principle of expected utility maximization. The antecedent of the right-hand side contains a condition on knowledge of the utility structure and on probabilistic beliefs about what player i himself and his opponent j will play. In the consequent it is stated that i will maximize his expected utility given his knowledge and beliefs.

As announced, additional axioms are needed to fix the belief formation policies of the players. First, players do not revise their beliefs during game play as long as this does not lead to inconsistency:

StratPers $\quad \bigwedge_i \bigwedge_j \mathbf{P}_i^x(j_k^z) = \mathbf{P}_i^y(j_k^z)$ for $x \preceq y \preceq z$.

More precisely, this belief persistence axiom states that if $x \preceq y \preceq z$, then the beliefs that player i has at x about the action of his opponent or himself at z will be the same at y. Of course, if game play has passed z and the beliefs have been contradicted, then i will have different beliefs. But as long as z has not been reached the beliefs remain constant.

StratPers concerns beliefs about strategies; RatPers, beliefs about rationality. It states that a player will never give up his beliefs in someone's rationality as long as that person has not moved; in more technical vocabulary, that if i believes at x that j is rational at some future node y, then i will not change that belief as long as j has not moved:

RatPers $\quad \bigwedge_i \bigwedge_j (\mathbf{K}_i^x \mathsf{Rrat}_j^y \leftrightarrow \mathbf{K}_i^x \mathsf{Rrat}_j^z)$ where $x \preceq y \prec z$, $\iota(z) = j$, and no u with $\iota(u) = j$ exists such that $y \prec u \prec z$.

It is left to the reader to verify that a striking consequence of this is that either i believes j to be rational everywhere, or nowhere.

4.2 Inconsistency result Given an extensive game with perfect information Γ, let proof system ${}_\Gamma\mathbf{KD_C P}\mathsf{Rrat}$ consist of the following axioms: all propositional tautologies, K, D, C, all axioms for linear (in)equalities and probabilistic reasoning, the proof rules modus ponens, necessitation, and induction, all axioms for many-moment game-playing situations for Γ, plus the three rationality and persistence axioms given above. Since our rendering does not use the T axiom, the epistemic assumption is better described as common true belief about rationality.

Theorem 4.1 *There is an extensive game with perfect information such that for all game-playing situations that consist of at least two decision moments there cannot be common knowledge of rationality at the second decision moment.*

Reny's original proof involves showing that no game-playing situation of the game shown in Figure 1 can have common knowledge of rationality at its second moment: every second decision moment would be a moment at which at x_1 player 2 has to move. But as in this game there is no way for both players to play on and gain (only one will gain from playing on), the suggestion may arise that at the end the inconsistency result is not very surprising. However, the result holds for games where both players would gain from playing on, too, and to underscore this we use in the proof the game shown in Figure 2 rather than Reny's original game shown in Figure 1. In fact, Reny's [35] characterization result reveals that the class of games for which inconsistency results can be proved is fairly large.

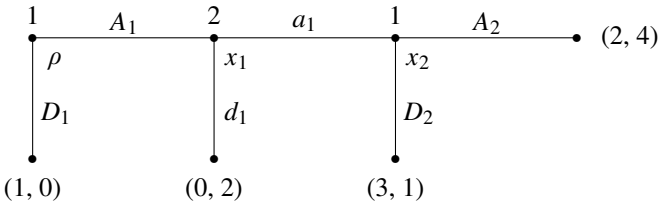

Figure 2 A Centipede Game

Proof We can make use of the two interrelation axioms Cons and KnProb because all relevant beliefs in this proof are probability one beliefs.

Claim 1 $\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to \mathbf{K}_2^{x_1}\mathbf{K}_1^{\rho}d_1$.

Proof of Claim 1 It suffices to show

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to \mathbf{K}_1^{\rho}d_1, \tag{1}$$

because from this a simple argument using the rule of necessitation for $\mathbf{K}_2^{x_1}$ would finish the proof. Because of the StratPers axiom, however, to prove (1) it suffices to show

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to \mathbf{K}_1^{x_1}d_1. \tag{2}$$

To show (2), in turn, we prove

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to (\mathbf{K}_1^{x_1}\mathsf{Rrat}_2^{x_1} \wedge \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}D_2), \tag{3}$$

and then apply necessitation for $\mathbf{K}_1^{x_1}$ to an instance of the rationality axiom to get

$$\vdash (\mathbf{K}_1^{x_1}\mathsf{Rrat}_2^{x_1} \wedge \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}D_2) \to \mathbf{K}_1^{x_1}d_1$$

to finish the proof of Claim 1. The remainder of the proof of Claim 1 is devoted, then, to showing (3). Clearly we have

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to \mathbf{K}_1^{x_1}\mathsf{Rrat}_2^{x_1}.$$

To prove

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}D_2, \tag{4}$$

observe that with the RatPers axiom for $\mathbf{K}_2^{x_1}$ and necessitation for $\mathbf{K}_1^{x_1}$ it can be shown that

$$\vdash \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_1} \to \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_2},$$

because x_2 is a successor of x_1 at which 1 moves for which in addition no y with $x_1 \succ y \succ x_2$ exists at which it is 2's turn. Hence

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_2}.$$

Applying the rationality axioms yields (4) concluding the proof of Claim 1.

Claim 2 $\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \to \mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{\rho}$.

Proof of Claim 2 Easy consequence of the RatPers axiom.

Claim 3 $\vdash (\mathbf{K}_2^{x_1}\mathbf{K}_1^{\rho}d_1 \wedge \mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{\rho}) \rightarrow \mathbf{K}_2^{x_1}\neg A_1$.

Proof of Claim 3 This is straightforward from the rationality axioms plus appropriate application of the rule of necessitation.

Claim 4 $\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \rightarrow \bot$.

Proof of Claim 4 Because of the KnWhere axiom we have $\vdash \mathbf{K}_2^{x_1}A_1$. Now we show that

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \rightarrow \neg\mathbf{K}_2^{x_1}A_1$$

to complete the proof. Combining Claim 1 and Claim 2 gives

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \rightarrow (\mathbf{K}_2^{x_1}\mathbf{K}_1^{\rho}d_1 \wedge \mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{\rho})$$

to which application of Claim 3 gives

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \rightarrow \mathbf{K}_2^{x_1}\neg A_1.$$

An application of the D axiom finishes the proof of Claim 4, and of the theorem. □

4.3 Discussion Given the game theoretic view of rationality as expected utility maximization, we should answer not so much the question whether RRat is plausible, but rather whether the belief persistence principles embodied in StratPers and RatPers are plausible. We distinguish the plausibility of the principles in general, and the plausibility of the specific instances in the proof of Theorem 4.1.

The general plausibility of the StratPers axiom first; that is,

$$\bigwedge_i \bigwedge_j \mathbf{P}_i^x(j_k^z) = \mathbf{P}_i^y(j_k^z)$$

for $x \preceq y \preceq z$. A possible argument in favor of this principle would be this. If at x player i believes that at some $z \succeq x$ his opponent j will choose action a, say, then there is no need for i to revise his beliefs at some intermediate y (satisfying $x \preceq y \preceq z$, that is) as long as i has not received any contradictory information on his way from x to y. But information contradicting that j choose a can only be information that j chooses, at z, an action different from a. Such information i cannot have received at the intermediate y. So at y player i will not need to revise his beliefs. Arguably, this yields a defense of StratPers.

Yet this argument overlooks subtle ways of obtaining pertinent information. A reason for player i's belief that j will choose a at z may be his belief that at z player j will choose rationally. If, however, on the path from x to y, player i has seen j choosing irrationally, this reason is no longer available. Player i might revise his beliefs in such a way that at z player j will play irrationally, too, and not choose a.

This is a general problem with the StratPers axiom. It completely ignores the fact that the reasons that one may have for particular beliefs may change over time, and that consequently one will need to reconsider or even revise one's beliefs, even if they are not directly contradicted by observed facts.

Similar arguments work against the general plausibility of the RatPers axiom; that is,

$$\bigwedge_i \bigwedge_j (\mathbf{K}_i^x\mathsf{Rrat}_j^y \leftrightarrow \mathbf{K}_i^x\mathsf{Rrat}_j^z),$$

where $x \preceq y \prec z$, $\iota(z) = j$, and no u with $\iota(u) = j$ exists such that $y \prec u \prec z$. For imagine that only irrational play on the part of j may get him from y to z. Although i believes, at x, that j will not take that irrational route, it is still questionable whether i should maintain that even if j plays irrationally he will, at node z, return to playing rationally.

Now it may of course be that the use of the belief persistence axioms in the proof of Theorem 4.1 is harmless or unproblematic. Let us, then, see where they are used in the proof. StratPers is used to prove

$$\vdash \mathbf{K}_1^{x_1} d_1 \rightarrow \mathbf{K}_1^{\rho} d_1$$

in Claim 1. The general problem that confronts us is clearly revealed. The reasons for the belief $\mathbf{K}_1^{x_1} d_1$ are the beliefs $\mathbf{K}_1^{x_1}\mathsf{Rrat}_2^{x_1} \wedge \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1} D_2$. This is so because $\vdash (\mathbf{K}_1^{x_1}\mathsf{Rrat}_2^{x_1} \wedge \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1} D_2) \rightarrow \mathbf{K}_1^{x_1} d_1$ is obtained by necessitation on an instance of the RRat axiom. But these reasons, although perhaps available at x_1, may not be available at ρ. That is, it may be doubted whether $\mathbf{K}_1^{\rho}\mathsf{Rrat}_2^{x_1} \wedge \mathbf{K}_1^{\rho}\mathbf{K}_2^{x_1} D_2$.

One way to substantiate doubt would concern the second conjunct. As inspection of the proof of Claim 1 shows, the reasons for $\mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1} D_2$ involve inter alia (1's beliefs about 2's beliefs about) the rationality of player 1 at x_2, or $\mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_2}$. The question whether this may figure as reasons for the beliefs at the root of the game (for $\mathbf{K}_1^{\rho} d_1$) then boils down to the question whether these reasons were already available at the root of the game, that is, whether $\mathbf{K}_1^{\rho}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_2}$ follows from $\mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_2}$.

It may come as an anticlimax that there do not seem to be any serious problems here. It is about player 1 imagining (at the root and at x_1) what player 2 will or does believe at x_1 about player 1 at x_2. But player 1 will not have obtained any new information about player 2's beliefs (at x_1!) while going from the root to x_1. At the root, player 1 imagines player 2's beliefs at x_1. And at x_1, player 1 imagines player 2's beliefs then and there, once again. There is no difference between these cases. There would have been a difference had the statement compared player 2's beliefs at the root with his beliefs at x_1. But that is not the issue here. Conclusion: StratPers causes no harm to the plausibility of the assumptions of Theorem 4.1.

What about RatPers? It is, first, used to prove

$$\vdash \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_1} \rightarrow \mathbf{K}_1^{x_1}\mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{x_2}.$$

One may find this problematic because it involves the rationality of player 1 at a decision moment when he need not pick an action. But apart from that there do not seem to be reasons to doubt this line of reasoning. Player 1 does not move at x_1, so player 2, if he believes that 1 is rational at the decision moment corresponding to x_1, has no reason to say that 1 would not be rational at the possible succeeding decision moment. And player 1 believes all this. An anticlimax again: RatPers is unproblematic here. RatPers is, however, also used to prove

$$\vdash \mathbf{C}^{x_1}\mathsf{Rrat}^{x_1} \rightarrow \mathbf{K}_2^{x_1}\mathsf{Rrat}_1^{\rho}.$$

And here we find something dubious at the end, for it is here that a belief revision policy is forced upon player 2 that is rather excessively rigid. It excludes, for instance, sensible dealings with a situation of the following kind. Player 2 has actually arrived at x_1; so player 1 has moved across. Player 2 believes that this was irrational but he also thinks that it was only an incident or an accident or a mistake. He believes at x_2 that player 1 was irrational at the first decision moment, but he also believes at x_2 that player 1 is rational at the second decision moment (and perhaps even the

third). This kind of subtle belief revision policy is excluded by RatPers. Either a player is believed to be rational everywhere or irrational everywhere.

To summarize, the proof of Theorem 4.1 boils down to showing that there is a contradiction between having arrived at x_1 and there being common knowledge of rationality at x_1. Such a contradiction can only be shown if, from the fact that there is common knowledge of rationality at x_1, it can be derived that one cannot be at x_1: more specifically, that one cannot be at x_1 since it can only be reached irrationally. That can only be demonstrated successfully if, from common knowledge of rationality at x_1, something follows about the beliefs and rationality at ρ.

But there is nothing in the concept of common knowledge at some decision moment that forces us to interpret it in such a temporally extended way and to adopt corresponding belief revision policies. In other words, there is nothing against allowing a game playing situation in which at the first decision moment there is no common knowledge of rationality, while there is in the second. The RatPers axiom (together with the StratPers axiom) exclude that possibility. This means that they are too strict. As an aside, one may find fault with Reny's inconsistency result on the basis of the fact that it presupposes the KnStratM axiom. For, as we have seen, from this axiom it follows that players already know what action they will pick at every future decision node, thus obviating some of the point of the many-moment interpretation.

5 Conclusion

We have developed a logical system to capture two different interpretations of what extensive games model. We have applied this to a long-standing debate in game theory between those who defend the claim that common knowledge of rationality leads to backward induction or the subgame perfect (Nash) equilibrium, and those who reject the claim. In particular, the logical analysis reveals that a defense of the claim à la Aumann [2] holds on to a conception of extensive game playing as a one-shot event in combination with a principle of rationality that is incompatible with it, while a rejection of the claim à la Reny [34] assumes a temporally extended, many-moment interpretation of extensive games in combination with implausible belief revision policies.

Apart from offering two interpretations of extensive game play, the logical machinery devised to analyze the defense of the claim à la Aumann was seen to be interesting in itself as it provides an inductive and implicit axiomatization of rationality in extensive games based on relations of dominance rather than the usual direct axiomatization of rationality as maximization of expected utility.

Some open questions remain. First, how general is the framework we present? Although a full answer cannot be given at this stage, we have shown elsewhere that several well-known, normal form game characterization theorems can be treated in our framework (De Bruin [17]). Furthermore, in a paper criticizing a characterization result due to Dekel and Fudenberg [21] we have proved a new characterization result for normal form games (De Bruin [18]). This shows the framework to be fruitful and flexible.

Second, can the formal system be extended to cover epistemic characterizations of solution concepts for extensive games with imperfect information? This will be more difficult because many such solution concepts are already peculiarly epistemic

in themselves, which makes it hard to separate epistemic conditions from solution concept.

Third and finally, do the present observations bear any relevance to the backward induction paradox (Sorensen [37], and references therein)? We believe they do, especially insofar as it is suggested that a solution to the paradox should distinguish one-shot interpretations from many-moment ones. This, however, is at present still work for the future.

References

[1] Asheim, G. B., and M. Dufwenberg, "Admissibility and common belief," *Games and Economic Behavior*, vol. 42 (2003), pp. 208–34. Zbl 1052.91007. MR 1984244. 270

[2] Aumann, R. J., "Backward induction and common knowledge of rationality," *Games and Economic Behavior*, vol. 8 (1995), pp. 6–19. Nobel Symposium on Game Theory (Björkborn, 1993). Zbl 0833.90132. MR 1315988. 261, 262, 265, 266, 277

[3] Aumann, R. J., "Reply to Binmore," *Games and Economic Behavior*, vol. 17 (1996), pp. 138–46. 261

[4] Aumann, R. J., "On the centipede game," *Games and Economic Behavior*, vol. 23 (1998), pp. 97–105. Zbl 0911.90354. MR 1618941. 261

[5] Aumann, R. J., "Interactive epistemology. I. Knowledge," *International Journal of Game Theory*, vol. 28 (1999), pp. 263–300. MR 1711434. 262

[6] Baltag, A., "A logic for suspicious players: Epistemic actions and belief-updates in games," *Bulletin of Economic Research*, vol. 54 (2002), pp. 1–45. MR 1883623. 262

[7] Barwise, J., "An introduction to first-order logic," pp. 5–46 in *Handbook of Mathematical Logic*, edited by J. Barwise, Elsevier, Amsterdam, 1977. 264

[8] Basu, K., "Strategic irrationality in extensive games," *Mathematical Social Sciences*, vol. 15 (1988), pp. 247–60. Zbl 0658.90110. MR 947868. 261

[9] Ben-Porath, E., "Rationality, Nash equilibrium and backwards induction in perfect-information games," *Review of Economic Studies*, vol. 64 (1997), pp. 23–46. Zbl 0890.90184. MR 1433544. 261

[10] van Benthem, J., "Games in dynamic-epistemic logic," *Bulletin of Economic Research*, vol. 53 (2001), pp. 219–48. MR 1857518. 262

[11] Bicchieri, C., "Self-refuting theories of strategic interaction: A paradox of common knowledge," *Erkenntnis*, vol. 30 (1989), pp. 69–85. 261

[12] Binmore, K., "Rationality and backward induction," *Journal of Economic Methodology*, vol. 4 (1997), pp. 23–41. 261

[13] Bonanno, G., "Modal logic and game theory: Two alternative approaches," *Risk, Decision, and Policy*, vol. 7 (2002), pp. 309–24. 262

[14] Bonanno, G., "A syntactic characterization of perfect recall in extensive games," *Research in Economics*, vol. 57 (2003), pp. 201–17. 262

[15] Brandenburger, A., A. Friedenberg, and H. Keisler, "Admissibility in games," *Econometrica*, vol. 76 (2008), pp. 307–52. 270

[16] Broome, J., and W. Rabinowicz, "Backwards induction in the centipede game," *Analysis*, vol. 59 (1999), pp. 237–42. Zbl 0943.91502. MR 1742621. 261

[17] de Bruin, B., *Explaining Games: On the Logic of Game Theoretic Explanations*, Diss., University of Amsterdam, Amsterdam, 2004. 277

[18] de Bruin, B., "Common knowledge of payoff uncertainty in games," *Synthese*, vol. 163 (2008), pp. 79–97. 270, 277

[19] Camerer, C., *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, Princeton, 2003. Zbl 1019.91001. 267

[20] Clausing, T., "Doxastic conditions for backward induction," *Theory and Decision*, vol. 54 (2003), pp. 315–36. Zbl 1069.91013. MR 2023457. 261

[21] Dekel, E., and D. Fudenberg, "Rational behavior with payoff uncertainty," *Journal of Economic Theory*, vol. 52 (1990), pp. 243–67. Zbl 0721.90084. MR 1082684. 270, 277

[22] Fagin, R., and J. Y. Halpern, "Reasoning about knowledge and probability," *Journal of the Association for Computing Machinery*, vol. 41 (1994), pp. 340–67. Zbl 0806.68098. MR 1369203. 264

[23] Feinberg, Y., "Subjective reasoning—Dynamic games," *Games and Economic Behavior*, vol. 52 (2005), pp. 54–93. Zbl 1099.91024. MR 2145701. 262

[24] Feinberg, Y., "Subjective reasoning—Solutions," *Games and Economic Behavior*, vol. 52 (2005), pp. 94–132. Zbl 1099.91025. MR 2145702. 262

[25] Fudenberg, D., and J. Tirole, *Game Theory*, The MIT Press, Cambridge, 1991. MR 1124618. 265, 270

[26] Halpern, J. Y., "Substantive rationality and backward induction," *Games and Economic Behavior*, vol. 37 (2001), pp. 425–35. Zbl 1027.91011. MR 1866185. 262

[27] Kaneko, M., "Epistemic logics and their game theoretic applications: Introduction," *Economic Theory*, vol. 19 (2002), pp. 7–62. Zbl 0995.03014. MR 1878957. 262

[28] Moulin, H., *Game Theory for the Social Sciences. Studies in Game Theory and Mathematical Economics*, New York University Press, New York, 1982. Zbl 0626.90095. MR 663188. 270

[29] Osborne, M. J., and A. Rubinstein, *A Course in Game Theory*, The MIT Press, Cambridge, 1994/1998. Zbl 0789.90092. MR 1301776. 265

[30] Pauly, M., "A modal logic for coalitional power in games," *Journal of Logic and Computation*, vol. 12 (2002), pp. 149–66. 262

[31] Pauly, M., "Programming and verifying subgame-perfect mechanisms," *Journal of Logic and Computation*, vol. 15 (2005), pp. 295–316. Zbl 1101.68685. MR 2195045. 262

[32] Pietarinen, A.-V., "Propositional logic of imperfect information: Foundations and applications," *Notre Dame Journal of Formal Logic*, vol. 42 (2001), pp. 193–210 (2003). Zbl 1034.03035. MR 2010181. 262

[33] Rabinowicz, W., "Grappling with the centipede: Defence of backward induction for BI-terminating games," *Economics and Philosophy*, vol. 14 (1998), pp. 95–126. 261

[34] Reny, P. J., "Common knowledge and games with perfect information," pp. 363–69 in *Philosophy of Science Association 1988, Vol. 2*, 1988. 261, 262, 271, 277

[35] Reny, P. J., "Common belief and the theory of games with perfect information," *Journal of Economic Theory*, vol. 59 (1993), pp. 257–74. Zbl 0802.90126. MR 1215147. 261, 273

[36] Samuelson, L., "Dominated strategies and common knowledge," *Games and Economic Behavior*, vol. 4 (1992), pp. 284–313. Zbl 0749.90092. MR 1155708. 270

[37] Sorensen, R., "Paradoxes of rationality," pp. 257–77 in *The Handbook of Rationality*, edited by A. Mele, Oxford University Press, Oxford, 2004. 278

[38] Stalnaker, R., "Knowledge, belief and counterfactual reasoning in games," *Economics and Philosophy*, vol. 12 (1996), pp. 133–63. 262

[39] Stalnaker, R., "Extensive and strategic forms: Games and models for games," *Research in Economics*, vol. 53 (1999), pp. 293–319. 261

[40] von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1944. Zbl 1112.91002. MR 0011937. 265

[41] Wolter, F., "First order common knowledge logics," *Studia Logica*, vol. 65 (2000), pp. 249–71. Zbl 0963.03024. MR 1775430. 262

Acknowledgments

Warmest thanks are due to Johan van Benthem, Christina Bicchieri, Peter van Emde Boas, Paul Harrenstein, Wiebe van der Hoek, Philip Reny, Martin Stokhof, and an anonymous referee of this journal for detailed comments on earlier versions of this paper.

Faculty of Philosophy
University of Groningen
Oude Boteringestraat 52
9712 GL Groningen
THE NETHERLANDS
b.p.de.bruin@rug.nl
http://www.philos.rug.nl/~debruin